nexenta
Enterprise class storage
for everyone

Data Sheet

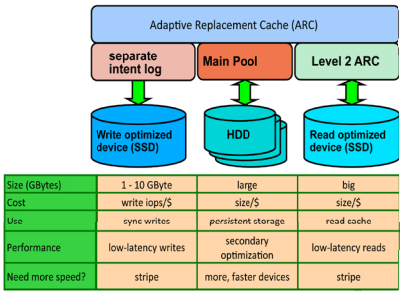
Auto-Tiered Storage

Conquering I/O Bottlenecks with Hybrid Storage

Overview

For more than a decade, storage system performance has remained rather stagnant while drive capacities and application performance demands steadily have increased. The result of this trend is an expensive problem: Storage users are forced into buying expensive hard disk drives (HDDs) to get a moderate performance boost (by reducing I/O latency) and/or forced into over-buying capacity in order to meet performance requirements.

With the advent and decreasing price of flash, storage vendors are integrating it into their products to solve this problem. ZFS technology is leading the industry in its ability to automatically and intelligently use flash in a storage system that offers the appropriate capacity and performance capabilities at a total cost that is dramatically lower than most legacy storage systems.



Hybrid Storage Pools

The ZFS Hybrid Storage Pool

ZFS is a robust, scalable file system with features not available in other file systems available today. One of these revolutionary features is the ZFS Hybrid Storage Pool (HSP), which allows you to combine DRAM, SSDs, and spinning HDDs into an accelerated storage medium. Below we will explore each of these offerings.

Adaptive Replacement Cache (ARC): The ARC lives in DRAM. It is the first destination for all data written to a ZFS pool, and it is the fastest (lowest-latency) source for data read from a ZFS pool.

When data is requested from ZFS, it first looks to the ARC; if it is there, it can be retrieved extremely fast (typically in nanoseconds) and provided back to the application. The contents of the ARC are balanced between the most recently used (MRU) and most frequently used (MFU) data.

Level-Two ARC (L2ARC): The L2ARC lives in SSDs. In concept, it is an extension of the ARC. Without an L2ARC, data that could not fit in the ARC would have to be retrieved from HDDs when requested. That is when drive speed makes a difference, but the performance difference between “fast” (e.g., 15k-RPM) and “slow” (e.g., 7,200-RPM) is in terms of latencies measured as a few milliseconds or several milliseconds; both are dramatically slower than ARC accesses measured in nanoseconds. L2ARC, in SSDs, fits nicely between the two—both in terms of price and performance.

Buying hundreds of gigabytes of flash is cheaper than the same capacity of DRAM (though still more expensive today than HDDs), and flash’s I/O latencies typically are measured in microseconds—slower than DRAM but still far faster than even “high-performance” HDDs. The L2ARC is populated by data first placed in the ARC as it becomes apparent that the data might get squeezed out of the ARC, and not every piece of data that existed in ARC will make it to the L2ARC (those that do not would be retrieved from HDDs instead, if requested); the algorithms that manage L2ARC population are automatic and intelligent.

ZFS Intent Log (ZIL): The ZIL is used to handle synchronous writes—write operations that are required by protocol (e.g., NFS, SMB/CIFS) to be stored in a non-volatile location on the storage device before they can be acknowledged—to ensure data stability. Databases usually require transactions to be on stable storage devices before continuing, so reducing the latency of synchronous writes has a direct impact on performance.

ZFS can do this by using placing the ZIL on a SSD. All writes (whether synchronous or asynchronous) are written into the ARC in DRAM, and synchronous writes are also written to the ZIL before being acknowledged.

Under normal conditions, ZFS regularly bundles up all of the recent writes in the ARC and flushes them to the spinning drives—at which point the data in the ZIL is no longer relevant (because it now exists on its long-term, non-volatile destination) and can be replaced by new writes. The ZIL only is read from when synchronous writes in the ARC are unable to be written to spinning disk—like after a power failure or controller failover—at which point ZFS reads the ZIL and places that data onto the spinning drives as intended. One might compare this concept to non-volatile RAM (NVRAM) from storage vendors like NetApp, but where NVRAM uses batteries that can wear out and have other issues, write-optimized SLC (single-level cell) flash devices do not need batteries.

And while NVRAM scalability is limited to available slots, adding SLOGs is as easy as adding HDDs. Like L2ARC, the ZIL/SLOG is managed automatically and intelligently by ZFS: Writes that need it are accelerated, without any additional effort by the administrator.

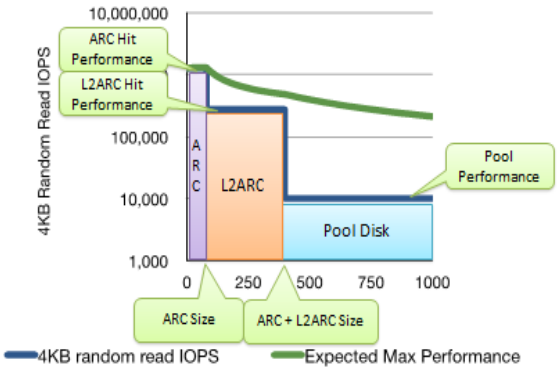
Hard Disk Drives (HDD): With the ARC, L2ARC, and ZIL/SLOG providing the bulk of the performance from a ZFS Hybrid Storage Pool, spinning drives are relegated to the job they do well—providing lower-performance, higher-density, low-cost storage capacity. Until the day that flash competes with HDDs on a dollar-per-gigabyte front, the right balance of DRAM and flash for performance, and HDDs for capacity, results in a total cost of ownership (TCO) that is less—both initially and over the long-term—than solving both requirements using all flash or all HDDs.

A New Storage Parameter: Working Set Size

For legacy storage systems, sizing means determining necessary capacity, IOPS, and throughput and then performing some simple math to determine the number of spindles that could provide those numbers.

As the industry moves towards more sophisticated caching methodologies in storage systems, a new parameter for expressing storage needs has become evident: The Working Set Size (WSS) can be described as the subset of total data that is actively worked upon (e.g., 500GB of this quarter's sales data out of a total database of 20TB).

Knowing the WSS makes it possible to size ARC, L2ARC, and even HDDs more accurately, but few applications today have an awareness of WSS.



Working Set Size

Conclusion

ZFS hybrid storage pools intelligently combine DRAM, flash, and hard disk drives to achieve the right balance of cost and performance for any given working set, while reducing the need for administrators to constantly monitor storage for I/O bottlenecks.

By reducing both read and write latency with the use of flash in a ZFS hybrid storage pool, we end up with a system that performs far better than legacy storage systems, while having a much lower total cost of ownership (TCO).

nexenta
Enterprise class storage for everyone

Nexenta Systems is the leading supplier of enterprise-class OpenStorage solutions. Its flagship software-only platform, NexentaStor, delivers high-performance, ultra-scalable, cloud- and virtualization-optimized storage solutions.

Nexenta Systems, Inc.
455 El Camino Real
Santa Clara, CA 95050
www.nexenta.com
www.facebook.com/nexenta
twitter.com/nexenta

Nexenta Systems EMEA
Camerastraat 8,
1322 BC, Almere
The Netherlands